R Practice Assignment : Week 5

Mohammed Saif Wasay (002815958)

MPS Informatics, Northeastern University

ALY 6010: Probability Theory and Introductory Statistics

Harpreet Sharma

June 22th, 2024

Introduction

We will analyze a student performance dataset in this practice assignment. The dataset is made up of six columns that represent a student's life style and performance. The data includes the number of hours a student spent studying, their prior score, whether they participate in any extracurricular activities, their sleep hours, the number of sample papers they have completed, and lastly their performance index. The performance index is a dependent variable in this case, while the others are independent variables. To put this to the test, we create a regression model.

To begin our analysis, we load the data into R and format the columns, as well as check for null values. After that, the null values are eliminated from the dataset.

1. **Correlation between Performance Index and Hours Studied:**

- **Null Hypothesis (H0):** There is no significant correlation between the performance index and the number of hours studied by students. Mathematically, this can be stated as:

$$H0 : \rho=0$$

where ρ(rho) represents the population correlation coefficient.

- **Alternative Hypothesis (H1):** There is a significant positive correlation between the performance index and the number of hours studied by students. Mathematically, this can be stated as:

$$H1 : \rho\neq0$$

2. **Correlation between Performance Index and Hours Slept:**

- **Null Hypothesis (H0):** There is no significant correlation between the performance index and the number of hours slept by students. Mathematically, this can be stated as:

$$H0 : \rho=0$$

- **Alternative Hypothesis (H1):** There is a significant positive correlation between the performance index and the number of hours slept by students. Mathematically, this can be stated as:

$$H1 : \rho\neq0$$

3. **Correlation between Performance Index and Previous Scores:**

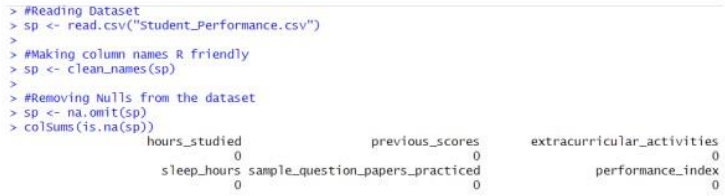
- **Null Hypothesis (H0):** There is no significant correlation between the performance index and the previous scores of students. Mathematically, this can be stated as:

$$H0 : \rho=0$$

- **Alternative Hypothesis (H1):** There is a significant positive correlation between the performance index and the previous scores of students. Mathematically, this can be stated as:

$$H1 : \rho\neq0$$

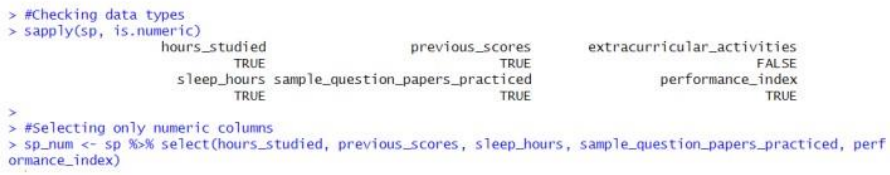
Figure 1: Data Cleaning



```
> #Reading Dataset
> sp <- read.csv("Student_Performance.csv")
>
> #Making column names R friendly
> sp <- clean_names(sp)
>
> #Removing Nulls from the dataset
> sp <- na.omit(sp)
> colSums(is.na(sp))
                  hours_studied                 previous_scores       extracurricular_activities
                              0                               0                                0
                    sleep_hours sample_question_papers_practiced                 performance_index
                              0                               0                                0
```

After cleaning the data, we examine the data kinds. We simply need the numeric variables to fit the data into a regression model. As a result, we build a subset of data that only contains numeric variables.

Figure 2: Checking Datatype



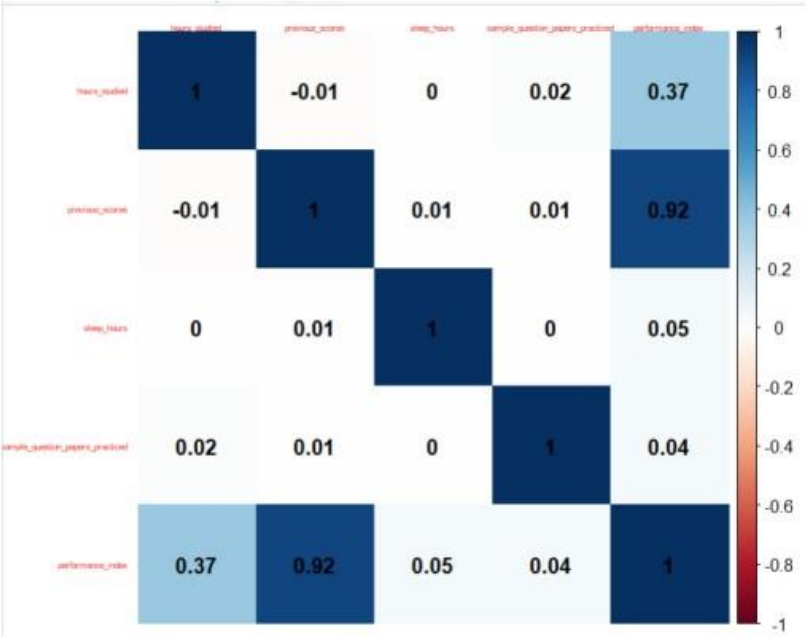
```
> #Checking data types
> sapply(sp, is.numeric)
                  hours_studied                 previous_scores       extracurricular_activities
                           TRUE                            TRUE                            FALSE
                    sleep_hours sample_question_papers_practiced                 performance_index
                           TRUE                            TRUE                             TRUE
>
> #Selecting only numeric columns
> sp_num <- sp %>% select(hours_studied, previous_scores, sleep_hours, sample_question_papers_practiced, perf
ormance_index)
```

To determine the pairwise association of our numeric variables sample_question_papers_practiced , hours_studied, sleep_hours, previous_scores, and performance_index, we generate a correlation matrix.

Figure 3: Correlation Matrix

```
> #Correlation Matrix for the data
> correlation_matrix <- cor(sp_num)
> correlation_matrix
                                 hours_studied previous_scores sleep_hours
hours_studied                      1.000000000    -0.012389916 0.001245198
previous_scores                   -0.012389916     1.000000000 0.005944219
sleep_hours                        0.001245198     0.005944219 1.000000000
sample_question_papers_practiced   0.017463168     0.007888025 0.003990220
performance_index                  0.373730351     0.915189141 0.048105835
                                 sample_question_papers_practiced performance_index
hours_studied                                         0.017463168        0.37373035
previous_scores                                       0.007888025        0.91518914
sleep_hours                                           0.003990220        0.04810584
sample_question_papers_practiced                      1.000000000        0.04326833
performance_index                                     0.043268327        1.00000000
```

Figure 4: Correlation Matrix heatmap



The correlation coefficient between the two relevant variables is represented by each cell in the matrix. The correlation coefficient is between -1 and 1. 1 denotes a perfect positive correlation (as one variable increases, so does the other), -1 denotes a perfect negative correlation (as one variable increases, so does the other), and 0 denotes no linear association.

All the variables will have a perfect correlation with themselves since it the same variable. The correlation matrix we can see that there is a very high positive correlation coefficient (0.92) between previous score of students and their performance index, and a positive moderate correlation (0.37) between performance index and hours a student studied. Finally, there is a small positive correlation between performance of students and the hours they slept and sample questions the solved.

Figure 5: Correlation Test for performance and hours studied

```
> #Correlation test on the performance and hours studied
> cor.test(sp_num$performance_index, sp_num$hours_studied, method = "pearson")

        Pearson's product-moment correlation

data:  sp_num$performance_index and sp_num$hours_studied
t = 40.289, df = 9998, p-value < 0.00000000000000022
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3567435 0.3904702
sample estimates:
      cor
0.3737304
```

We perform a correlation test between the performance index and hours studied, revealing a positive correlation coefficient of 0.3737. This suggests a positive linear relationship, implying that an increase in hours studied corresponds to an increase in the performance index. The p-value, remarkably small (below 0.00000000000000022), provides robust evidence against the null hypothesis, leading to the rejection of the null hypothesis in favor of the alternative. This signifies a significant correlation between the performance index and hours studied. The 95% confidence interval for the correlation coefficient, [0.3567435, 0.3904702], indicates a moderately strong positive correlation within this range.

Figure 6: Correlation Test for performance and hours slept

```
> #Correlation test on the performance and sleep_hours
> cor.test(sp_num$performance_index, sp_num$sleep_hours, method = "pearson")

        Pearson's product-moment correlation

data:  sp_num$performance_index and sp_num$sleep_hours
t = 4.8157, df = 9998, p-value = 0.000001489
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.02853267 0.06764213
sample estimates:
       cor
0.04810584
```

The positive correlation coefficient of 0.0481 suggests a weak positive linear relationship, indicating that higher values of performance_index are associated with higher sleep_hours values. The p-value, which is 0.000001489, is below the standard significance level of 0.05, providing strong evidence against the null hypothesis. With the small p-value, we reject the null hypothesis in favor of the alternative hypothesis, indicating a significant but very mild positive association between performance_index and sleep_hours. The 95% confidence interval for the correlation coefficient is [0.02853267, 0.06764213], signifying a range within which we can be 95% confident that the true correlation between the variables exists. In this context, it implies a minor positive association.

Figure 7: Correlation Test for performance and previous scores

```
> #Correlation test on the performance and previous_scores
> cor.test(sp_num$performance_index, sp_num$previous_scores, method = "pearson")

        Pearson's product-moment correlation

data:  sp_num$performance_index and sp_num$previous_scores
t = 227.06, df = 9998, p-value < 0.00000000000000022
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9119474 0.9183167
sample estimates:
      cor
0.9151891
```

A positive value (0.9152) indicates a very strong positive linear association, implying that when the variable previous_scores increases, so does the variable performance_index. The p-value is less than 0.00000000000000022 (which is extremely small). The p-value is extraordinarily low, indicating that there is substantial evidence to reject the null hypothesis. The true correlation is greater than zero is the alternate hypothesis. Given the very small p-value, the alternate hypothesis would be accepted and the null hypothesis rejected. This means that the performance_index and previous_scores have a significant and extremely strong positive association. The real correlation coefficient has a 95% confidence interval of [0.9119474, 0.9183167]. This interval specifies a set of values within which we can be 95% certain that the genuine correlation between the variables exists. It suggests an extraordinarily strong positive association in this example.

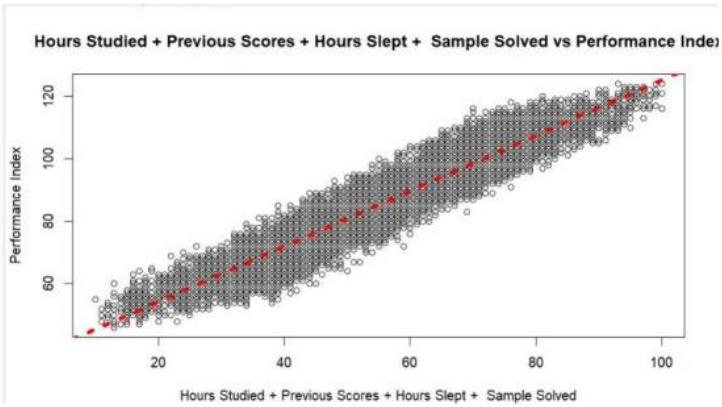
Figure 8: Fitting to predict performance index

```
> #Creating a fit for hours studies, sleep hours, sample question papers practiced, previous scores on perfor
mance index
> fit <-lm(hours_studied +  previous_scores + sleep_hours + sample_question_papers_practiced ~ performance_in
dex,data=sp_num)
> fit

Call:
lm(formula = hours_studied + previous_scores + sleep_hours +
    sample_question_papers_practiced ~ performance_index, data = sp_num)

Coefficients:
      (Intercept)  performance_index
          36.5564             0.8872

> summary(fit)

Call:
lm(formula = hours_studied + previous_scores + sleep_hours +
    sample_question_papers_practiced ~ performance_index, data = sp_num)

Residuals:
     Min       1Q   Median       3Q      Max
-16.9170  -3.8566   0.0004   3.7446  17.3388

Coefficients:
                  Estimate Std. Error t value            Pr(>|t|)
(Intercept)       36.556388   0.162102   225.5 <0.0000000000000002 ***
performance_index  0.887212   0.002772   320.0 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.326 on 9998 degrees of freedom
Multiple R-squared:  0.9111,    Adjusted R-squared:  0.9111
F-statistic: 1.024e+05 on 1 and 9998 DF,  p-value: < 0.00000000000000022
```

We are regressing a collection of independent variables on the dependent variable performance_index (hours_studied, sleep_hours, previous_scores, sample_question_papers_practiced). We can see from the summary of the fit that the expected intercept is 36.5564. When independent variables are zero, this is the expected value of performance_index. The performance_index coefficient is calculated to be 0.8872. Holding all other variables constant, this reflects the estimated change in the dependent variable for a one-unit change in performance_index. The p-values (Pr(>|t|)) are exceptionally small, showing strong evidence that the relevant coefficients are not equal to zero. The residuals are the disparities between the dependent variable's observed and expected values. The summary describes how they are distributed.

The residual standard error, a measure of the residuals' standard deviation, is approximately 5.326. The R-squared value is 0.9111, indicating that the model explains around 91.11% of the variability in the dependent variable. The adjusted R-squared, accounts for the total predictors, closely aligns with R-squared, suggesting minimal penalization for additional variables. The F-statistic is exceptionally high (1.024e+05), and the p-value is extremely small (0.0000000000000022), confirming the model's overall significance.

Figure 9: Scatter Plot of Independent variables and Performance Index



The fit line on the scatter plot indicates the link between the independent variable (performance_index) and the sum of the other independent variables (hours_studied, sleep_hours, previous_scores, sample_question_papers_practiced). The fit line reflects the prediction of the dependent variable (performance_index) by the linear regression model based on the sum of the other variables.

Conclusion

On the student performance dataset, we were able to clean the data, generate a correlation matrix, and plot a correlation heat map of our dependent and independent numeric variables. We were able to calculate the correlation coefficients and the correlation of independent factors on the students' performance index. We also ran correlation tests on the variables and analyzed the results. Finally, a regression model was developed to fit the dependent variable, the performance index. A higher percentage of R-square (91%) shows a better fit of the model, implying that the hours studied, previous scores, hours slept, and sample papers practiced explain a greater share of the variability in the performance index.

Citations

R Documentation, An introduction to R. Retrieved 22th June 2024 from https://cran.r-project.org/doc/manuals/r-release/R-intro.html#Related-software-and-documentation

Shaun Turney (2022). Retrieved 22th June 2024 Coefficient of Determination (R^2) Calculation & Interpretation https://www.scribbr.com/statistics/coefficient-of-determination/.